DOI 10.15276/imms.v13.no1-2.137
УДК 681.514:621.029

Informatics and Mathematical Methods in Simulation
Vol.13 (2023), No. 1-2, pp. 137-143

ВИЯВЛЕННЯ АУДІО-ПІДРОБОК ЗАСОБАМИ ШТУЧНОГО ІНТЕЛЕКТУ

М.А Стецовський., В.В.Зоріло, О.Ю.Лебедєва

Національний університет «Одеська політехніка»
пр.Шевченка, 1, Одеса, 65044
email: vikazorilo@gmail.com

Розвиток інформаційних технологій, зокрема, штучного інтелекту, призводить до широкого їх застосування у багатьох сферах нашого життя. Із стрімким розвитком штучного інтелекту зростає кількість випадків його застосування для генерації цифрових зображень, аудіо, відео тощо. Підробка цифрових аудіо файлів є небезпечною з точки зору використання зловмисниками для чинення інформаційно-психологічного впливу та маніпуляцій суспільством та окремими індивідами. Існують сучасні методи виявлення аудіо-підробок, виконаних засобами штучного інтелекту. Вони мають високу точність, при цьому не позбавлені недоліків. Основним недоліком є складна архітектура та висока ресурсоємність. Метою даною роботи є розробка нейронної мережі, яка дозволила б з задовільною точністю виявляти аудіо-підробку, та навчання якої не вимагало б значних обчислювальних ресурсів. Було обрано метод для модифікації, а саме метод виявлення аудіо підробок з використанням згорткової нейронної мережі. Було модифіковано метод виявлення аудіо підробок шляхом побудови моделі з новою архітектурою з меншою кількістю шарів, що дозволило значно скоротити часові витрати та потреби в значних обчислювальних ресурсах в порівнянні з аналогами. Отримані результати експериментів із застосуванням модифікованого методу показують задовільну ефективність і точність системи. Помилки 1 роду склали 24%, помилки другого роду – 9%. Розроблену модифікацію реалізовано у програмному додатку із зручним і простим інтерфейсом.
Ключові слова: штучний інтелект, нейронні мережі, аудіо-підробка, виявлення аудіо-фейку.

Вступ. З розвитком технологій поширення підробленого контенту викликає значне занепокоєння в різних сферах. У той час як візуальні підробки привертають значну увагу, аудіо підробки стали ще однією сферою, що викликає занепокоєння.

Аудіо підробки можуть завдати значної шкоди, включаючи поширення дезінформації, видавання себе за іншу особу та маніпуляції з аудіо доказами. Зі зростанням легкості створення переконливих аудіо підробок потреба в ефективних механізмах їх виявлення стає першочерговою.

Порівняно з візуальними підробками, кількість загальнодоступних наборів даних про аудіо підробки є відносно обмеженою. Такий дефіцит даних створює проблеми для навчання та оцінки моделей глибокого навчання, спеціально розроблених для виявлення аудіо підробок. Вирішення цієї проблеми вимагає спільних зусиль у створенні та поширенні різноманітних наборів даних про аудіо підробки.

Моделі глибокого навчання можуть вимагати значної обчислювальної потужності, що обмежує їхнє розгортання на пристроях з обмеженими ресурсами або платформах потокового мовлення в реальному часі. Оптимізація моделей для підвищення ефективності без втрати точності є ключовим напрямком досліджень.

Інтерпретованість моделей виявлення аудіо підробок є критично важливою проблемою. Розуміння та пояснення процесу прийняття рішень, що лежить в основі цих моделей, є важливим, особливо в юридичному або судовому контексті, де прозорість і

довіра мають першорядне значення. Розробка зрозумілих методів виявлення глибоких підробок є сферою активних досліджень.

Ефективними сучасними методами виявлення аудіо підробок є наступні.

DeepSpectrum – це модель глибокого навчання, спеціально розроблена для аналізу та класифікації спектроскопічних даних.

DeepSpectrum використовує можливості алгоритмів глибокого навчання для автоматичного вилучення значущих характеристик зі спектроскопічних даних і створення точних прогнозів або класифікацій. Навчаючись на великих масивах даних мічених спектрів, DeepSpectrum може вивчати складні закономірності та взаємозв'язки в даних, що дозволяє йому виконувати такі завдання, як ідентифікація матеріалів, контроль якості або виявлення аномалій [5].

Однією з ключових переваг DeepSpectrum є його здатність обробляти високорозмірні спектроскопічні дані, які часто містять численні спектральні смуги або канали. Модель може ефективно фіксувати і представляти складні спектральні особливості, що дозволяє проводити надійний аналіз і класифікацію.

Однак, як і будь-яка модель машинного навчання, DeepSpectrum також має певні обмеження. Для досягнення оптимальної продуктивності їй потрібна значна кількість маркованих навчальних даних, що може бути проблемою для областей з обмеженими або дефіцитними даними. Навчання моделі DeepSpectrum на наборі даних з 3000 вибірок потенційно може зайняти годину на епоху або навіть більше.

VGGish – це модель глибокого навчання, розроблена дослідницькою групою Google зі штучного інтелекту, спеціально призначена для завдань аналізу аудіо. Вона в першу чергу використовується для вилучення аудіо-вкладень або ознак зі спектрограм, які потім можуть бути використані для різних завдань, пов'язаних з аудіо, таких як класифікація аудіо, зіставлення схожості аудіо або виявлення аудіо-подій [6].

Однією з головних переваг VGGish є його здатність навчатися багатим і дискримінативним репрезентаціям з аудіо даних. Використовуючи свою глибоку архітектуру та велику кількість параметрів, що навчаються, VGGish може захоплювати як низькорівневі, так і високорівневі характеристики звуку, що робить його придатним для широкого спектру завдань аудіоаналізу.

Крім того, VGGish навчений на великому наборі аудіоданих, що дозволяє йому добре узагальнювати різні аудіодомени. Попередньо навчену модель VGGish, яка є загальнодоступною, можна використовувати як екстрактор ознак, де вихідні дані проміжних шарів моделі можуть бути використані як вбудовування аудіо для подальших завдань. Це дозволяє користувачам скористатися перевагами вивчених репрезентацій без необхідності тривалого навчання на власних наборах аудіоданих.

Ще однією перевагою VGGish є простота використання та інтеграції. Модель доступна як реалізація TensorFlow, що робить її сумісною з різними фреймворками глибокого навчання і дозволяє легко інтегрувати в існуючі робочі процеси.

Однак існує ризик надмірної адаптації, особливо при роботі з обмеженою кількістю мічених даних. Якщо навчальний набір даних відносно невеликий, модель може погано узагальнювати невидимі аудіо-зразки, що призведе до зниження продуктивності.

Навчання та використання моделі може вимагати значних обчислювальних ресурсів, включаючи пам'ять та обчислювальну потужність. Навчання моделі на низькопродуктивних або обмежених в ресурсах пристроях може бути складним, а висновок в реальному часі на таких пристроях також може бути складним в обчислювальному плані. Навчання моделі VGGish на наборі даних з 3000 вибірок потенційно може зайняти кілька годин на епоху або навіть більше.

Зазначені моделі дозволяють з високою точністю виявляти аудіо-підробки, однак це вимагає значних обчислювальних ресурсів, що ускладнює їх використання в побуті або в умовах неможливості доступу до інтернету або хмарних сервісів.

Метою даною роботи є розробка нейронної мережі, яка дозволила б з задовільною точністю виявляти аудіо-підробку, та навчання якої не вимагало б значних обчислювальних ресурсів.

Матеріали та методи. Як альтернативу зазначеним методам, основним недоліком яких є потреба у великих обчислювальних потужностях, в даній роботі запропоновано наступну модель нейронної мережі.

Модель, побудована за допомогою Sequential API, має шарувату структуру, яка поєднує шари Conv2D, MaxPooling2D, Flatten та Dense. Розглянемо кожен компонент більш детально:

Шари Conv2D: Використовуються два шари Conv2D з 16 фільтрами кожен. Ці шари виконують операції згортки над вхідною аудіо спектрограмою, яка представляє звук у візуальному форматі. Застосовуючи фільтри 3×3, модель може виокремлювати локальні особливості зі спектрограми, фіксуючи основні патерни, пов'язані з аудіо сигналом. Набір фільтрів, також відомих як ядра, у згортковому шарі згорткової нейронної мережі – це набір матриць, які використовуються для згортки з вхідними даними [1].

Кожне ядро виконує операцію згортки вхідних даних. Під час операції згортки ядро ковзає по вхідним даним (наприклад, по зображенню) і обчислює скалярний добуток між елементами ядра і відповідними вхідними елементами. Результатом цього обчислення є нова матриця, яка називається картою ознак або картою згортки.

Набір фільтрів у шарі згортки дозволяє виявляти різні локальні особливості вхідних даних. Кожен фільтр може спеціалізуватися на виявленні певних особливостей, таких як контури, текстури або форми. Кілька фільтрів можна використовувати для виявлення різних особливостей у різних частинах вхідних даних. Наприклад, деякі фільтри можуть виявляти вертикальні контури, інші – горизонтальні, а треті – особливості текстури.

Таким чином, набір фільтрів у згортковому шарі допомагає моделі виявляти і розпізнавати різні особливості у вхідних даних, формуючи карту особливостей, яка передається на наступний рівень мережі для подальшого аналізу і виконання завдань, таких як класифікація або виявлення об'єктів.

Функція активації: Функція активації Rectified Linear Unit (ReLU) використовується після кожного шару Conv2D. Функція ReLU (Rectified Linear Unit) є активаційною функцією, часто використовуваною в нейронних мережах. Вона приймає вхідне значення і повертає максимум між нулем і вхідним значенням. ReLU вносить нелінійність у модель, дозволяючи їй вивчати складні взаємозв'язки між виділеними ознаками. Ця нелінійна активація полегшує здатність моделі вловлювати складні деталі та дискримінаційні характеристики, присутні в аудіо даних. Математичне представлення ReLu функції наступне $f(x) = max(0, x)$.

Шари MaxPooling2D: Два шари MaxPooling2D слідують за шарами Conv2D. Ці шари виконують даунсемплінг, вибираючи максимальні значення у вікні об'єднання. Зменшуючи просторові розміри вхідних даних, MaxPooling2D допомагає зберегти найбільш важливу інформацію, зменшуючи при цьому обчислювальну складність [2].

Шар максимального пулінгу застосовується зазвичай після згорткових шарів у згорткових нейронних мережах. Його основна мета – зменшити просторові розміри карт ознак, ущільнюючи інформацію і витягуючи найважливіші ознаки.

Це допомагає знизити кількість параметрів у моделі, скоротити обчислювальну складність і зробити модель стійкішою до невеликих перекладацьких спотворень вхідних даних.

Зазвичай шари максимального пулінгу застосовуються послідовно із згортковими шарами для зменшення просторових розмірів карт ознак. Це дає змогу моделі зосередитися на найбільш значущих ознаках і покращує інваріантність до масштабування та переміщень об'єктів на зображенні.

Шар Flatten: Шар Flatten призначений для перетворення багатовимірних даних, отриманих на попередніх шарах, в одновимірний вектор. Це перетворення готує дані для наступних повністю з'єднаних шарів, дозволяючи моделі вивчати глобальні особливості звуку [3].

Після застосування згорткових шарів у згорткових нейронних мережах, вихідні дані можуть мати форму тривимірного тензора, наприклад, (batch_size, height, width, channels), де batch_size – розмір пакета (batch), height і width – розміри висоти і ширини, а channels – кількість каналів.

Операція Flatten() перетворює ці тривимірні дані в одновимірний вектор, об'єднуючи всі елементи в послідовність. Таким чином, кожен елемент тривимірного тензора стає окремим елементом одновимірного вектора.

Операція Flatten() використовується для переходу від згорткових шарів до повнозв'язаних шарів нейронної мережі. Після цієї операції дані можуть бути подані на шари з повнозв'язними нейронами, які очікують одновимірні входи. Це дає змогу нейронній мережі моделювати складні залежності між ознаками і зробити прогнози або класифікацію на основі цих даних.

Шар Dense:: Повнозв'язні шари (Dense Layers) у нейронних мережах відіграють роль об'єднання ознак, отриманих із попередніх шарів, і моделювання складніших залежностей у даних [4].

Вони встановлюють зв'язки між кожним нейроном у поточному шарі та кожним нейроном у попередньому шарі. Це означає, що кожен вихідний сигнал із попереднього шару впливає на кожен нейрон у повнозв'язному шарі.

Кожен нейрон у повнозв'язному шарі отримує зважену суму вхідних сигналів, де кожен вхідний сигнал множиться на свою відповідну вагу. Ваги визначаються під час навчання моделі і являють собою параметри, які модель намагається оптимізувати.

Після обчислення зваженої суми, нейрони повнозв'язного шару застосовують функцію активації до результату. Функція активації додає нелінійність у модель, даючи змогу моделі апроксимувати складні нелінійні залежності в даних.

Вихідний сигнал кожного нейрона в повнозв'язному шарі являє собою результат застосування функції активації до зваженої суми вхідних сигналів. Ці виходи можуть передаватися наступному шару в нейронній мережі або використовуватися для зробити прогнози, наприклад, у завданні класифікації.

Повнозв'язні шари дають змогу моделювати складніші залежності між ознаками та створювати більш гнучкі та виразні моделі. Їх часто використовують наприкінці нейронної мережі для ухвалення рішень або видачі остаточних виходів моделі.

До моделі включено два повнозв'язні шари. Перший щільний шар складається з 128 одиниць і використовує функцію активації ReLU. Цей шар виконує нелінійні перетворення вхідних даних, виділяючи ознаки вищого рівня. Останній щільний шар складається з одного елемента з сигмоїдною функцією активації, що дає ймовірнісний вихід для задач бінарної класифікації.

Вищезгадана архітектура моделі, представлена на рисунку 1, пропонує кілька помітних переваг для задач класифікації аудіо:

1) виділення локальних особливостей – шари Conv2D дуже ефективно виділяють локальні особливості в аудіо спектрограмі. Згортаючи фільтри над невеликими сприйнятливими полями, модель може вивчати значущі представлення локальних патернів, включаючи часові та спектральні характеристики. Ця здатність вловлювати дрібні деталі підвищує дискримінаційну здатність моделі;

2) ієрархічне представлення – шари Conv2D і MaxPooling2D, що накладаються один на одного, полегшують створення ієрархічного представлення звукових характеристик. Початкові шари захоплюють низькорівневі елементи, такі як краї, текстури та основні компоненти звуку. Коли інформація проходить через наступні шари, вивчаються більш складні та абстрактні представлення. Ця ієрархічна структура дозволяє моделі виокремлювати як низькорівневі, так і високорівневі ознаки з вхідної спектрограми;

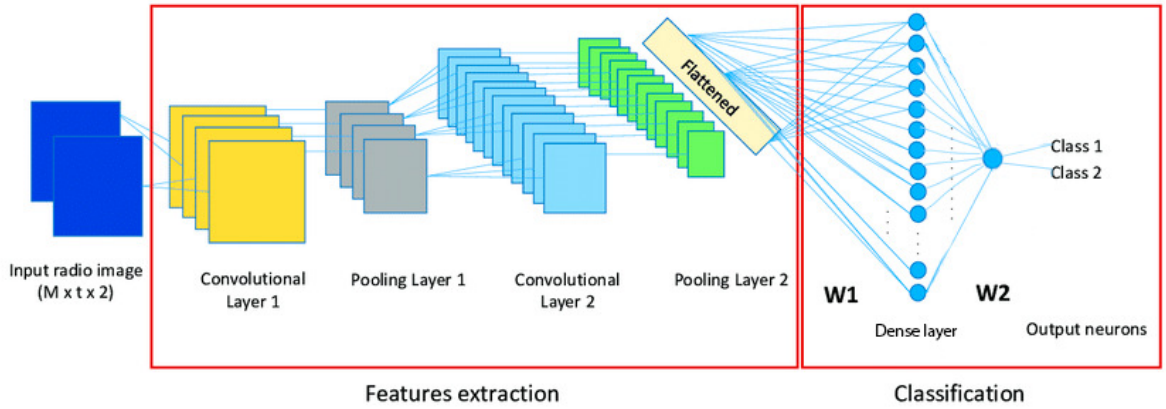
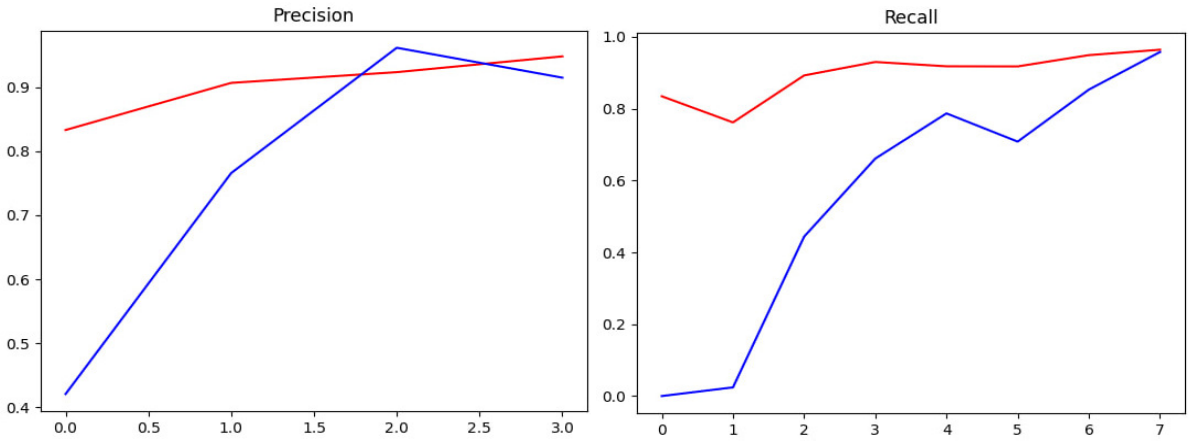


Рис. 1. Архітектура моделі

3) нелінійність і функції активації – включення функцій активації ReLU вносить в модель нелінійність. Ця нелінійність є життєво важливою для відображення складних взаємозв'язків, присутніх в аудіоданих, які часто демонструють нелінійні патерни і залежності. Фінальна сигмоїдна функція активації дозволяє моделі виробляти ймовірнісний вихід, що робить її придатною для задач бінарної класифікації;

4) ефективність параметрів: архітектура моделі забезпечує баланс між збором релевантної інформації та уникненням надмірного налаштування. Використовуючи помірну кількість фільтрів та операцій об'єднання, модель зменшує кількість параметрів порівняно з більш глибокими архітектурами. Така ефективність параметрів робить модель обчислювально ефективною, дозволяючи їй обробляти більші набори даних і зменшуючи ризик перенавчання, особливо при роботі з обмеженою навчальною вибіркою.

Результати. Тренування моделі на датасеті в 3000 екземплярів зайняло сумарний час в півгодини, було проведено 5 епох тренування. Виконано модифікацію методу виявлення аудіо підробок з використанням нейронної мережі. За попереднім результатом тренування впевненість склала близько 90%. Порівняно з аналогами це гірше відносно точності моделі, але краще в використанні обмеженої кількості ресурсів. На рис.2 червоним позначено результат, отриманий на навчальному датасеті, синім – на даних тестування.



а. б.

Рис. 2. Графіки впевненості (а) і повноти (б) моделі

Precision (впевненість) – це показник ефективності моделі в контексті бінарної класифікації, який вимірює частку вірно класифікованих позитивних результатів серед усіх екземплярів, які модель передбачила як позитивні. Він дає змогу оцінити, наскільки точно модель ідентифікує істинно позитивні екземпляри.

Recall (повнота) – це показник ефективності моделі, який вимірює здатність моделі правильно ідентифікувати позитивні екземпляри серед загальної кількості дійсних позитивних екземплярів у наборі даних. Він дає уявлення про те, наскільки модель здатна виявляти всі позитивні приклади або мінімізувати помилкові негативні результати.

Після проведення тестування зі 100 реальними та 100 підробленими екземплярами була визначена кількість помилок 1 та 2 роду (рис. 3).

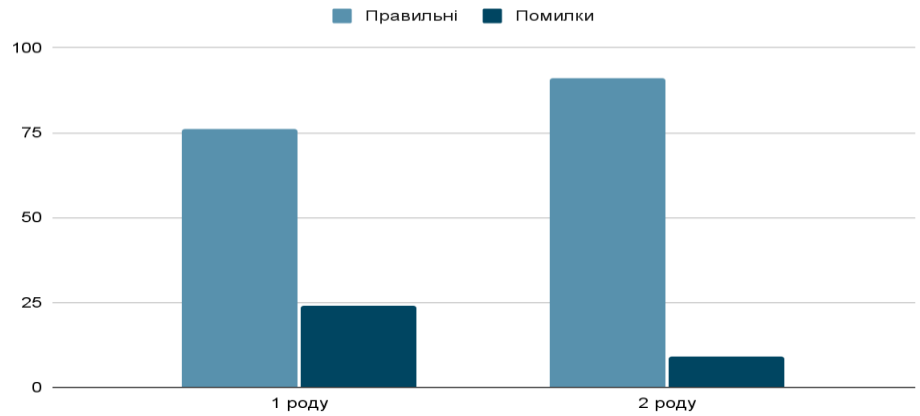


Рис.3. Помилки 1 та 2 роду

Помилки 1 роду склали 24%, 2 роду – 9% Загалом програма продемонструвала задовільну роботу та високу ефективність.

Помилки першого роду, або хибні позитивні спрацьовування, можуть іноді виникати в процесі аналізу даних або прийняття рішень. Вони можуть бути спричинені різними факторами, такими як статистична варіація або шум у даних. Однак, незважаючи на це, загальна кількість таких помилок залишається малою і не перевищує прийнятний рівень. Важливо зазначити, що програма все одно демонструє високу точність і надійність у виконанні своїх основних функцій.

Крім того, важливо усвідомлювати, що під час оцінювання ефективності програми необхідно враховувати й інші параметри, такі як показники точності, повноти та загальну здатність моделі передбачати правильні результати.

Список літератури

1. What is a convolutional layer?. Databricks. URL: https://www.databricks.com/glossary/convolutional-layer.
2. Papers with code – max pooling explained. The latest in Machine Learning | Papers With Code. URL: https://paperswithcode.com/method/max-pooling.
3. Layer – flatten. TensorSpace.js. URL: https://tensorspace.org/html/docs/layerFlatten.html.
4. Kaplan D. Dense layer: the building block to neural networks. URL: https://enjoymachinelearning.com/blog/dense-layer/.
5. GitHub – DeepSpectrum/DeepSpectrum. GitHub. URL: https://github.com/DeepSpectrum/DeepSpectrum.
6. VGGish. GitHub. URL: https://github.com/tensorflow/models/tree/master/research/audioset/vggish.

DETECTION OF AUDIO FAKES BY MEANS OF ARTIFICIAL INTELLIGENCE

M.A. Stetsovskyi, V.V.Zorilo, O.Yu. Lebedeva

National Odesa Polytechnic University,
1, Shevchenko Ave, Odesa, Ukraine
vikazorilo@gmail.com

The development of information technologies, including artificial intelligence, is leading to their widespread use in many areas of our lives. With the rapid development of artificial intelligence, the number of cases of its application to generate digital images, audio, video, etc. is growing. Counterfeiting digital audio files is dangerous from the point of view of being used by criminals to exert information and psychological influence and manipulate society and individuals. There are modern methods for detecting audio fakes made by artificial intelligence. They are highly accurate, but not without drawbacks. The main drawback is the complex architecture and high resource intensity. The aim of this paper is to develop a neural network that would allow detecting audio fakes with satisfactory accuracy and whose training would not require significant computing resources. A method was chosen for modification, namely, the method of detecting audio fakes using a convolutional neural network. The method for detecting audio fakes was modified by building a model with a new architecture with fewer layers, which significantly reduced the time and computational resources required compared to analogues. The experimental results obtained using the modified method show satisfactory efficiency and accuracy of the system. Errors of the first kind amounted to 24%, errors of the second kind - 9%. The developed modification is implemented in a software application with a convenient and simple interface.

Keywords: artificial intelligence, neural networks, audio forgery, audio fake detection